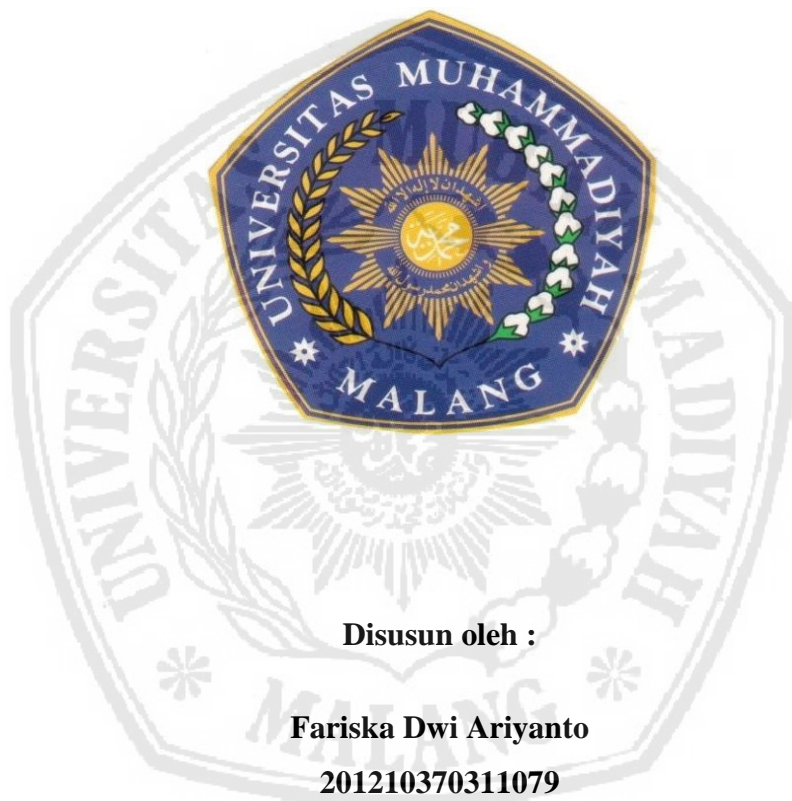


# **PERBANDINGAN METODE TF-IDF DAN TW-IDF DALAM PENCARIAN DOKUMEN TUGAS AKHIR UNIVERSITAS MUHAMMADIYAH MALANG**

## **TUGAS AKHIR**

**Diajukan Untuk Memenuhi  
Persyaratan Guna Meraih Gelar Sarjana Strata 1  
Teknik Informatika Universitas Muhammadiyah Malang**



**Disusun oleh :**

**Fariska Dwi Ariyanto**

**201210370311079**

**PROGRAM STUDI TEKNIK INFORMATIKA  
FAKULTAS TEKNIK  
UNIVERSITAS MUHAMMADIYAH MALANG  
2017**

**LEMBAR PENGESAHAN**  
**PERBANDINGAN METODE TF-IDF DAN TW-IDF DALAM**  
**PENCARIAN DOKUMEN TUGAS AKHIR UNIVERSITAS**  
**MUHAMMADIYAH MALANG**

**TUGAS AKHIR**

Sebagai Persyaratan Guna meraih Gelar Sarjana Strata 1  
Teknik Informatika Universitas Muhammadiyah Malang

Disusun Oleh :

**FARISKA DWI ARIYANTO**

**(201110370311079)**

Tugas Akhir ini telah diuji dan dinyatakan lulus melalui sidang majelis penguji  
pada tanggal 25 Januari 2017

Menyetujui,

Penguji I,



**Nur Hayatin, S.ST., M.Kom.**

**NIDN: 0726038402**

Penguji II,



**Galih Wasis W, S.Kom., M.Cs.**

**NIDN: 0723028801**

Mengetahui,

Ketua Jurusan Teknik Informatika



**Yuda Murni, S.Kom., M.Sc**

**NIDN: 0706077902**

## KATA PENGANTAR

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

Segala puji bagi Allah SWT, yang telah memberikan Rahmat dan Karunianya, sehingga penulis dapat menyelesaikan skripsi yang berjudul:

**“PERBANDINGAN METODE TF-IDF DAN TW-IDF DALAM  
PENCARIAN DOKUMEN TUGAS AKHIR UNIVERSITAS  
MUHAMMADIYAH MALANG ”**

Skripsi ini merupakan salah satu syarat studi yang harus ditempuh oleh seluruh mahasiswa Universitas Muhammadiyah Malang, guna menyelesaikan akhir studi pada jenjang program Strata 1.

Peneliti menyadari masih banyak kekurangan dan keterbatasan dalam penulisan tugas akhir ini. Untuk itu, penulis sangat mengharapkan saran yang membangun agar tulisan ini dapat berguna untuk perkembangan ilmu pengetahuan kedepan.

Malang, 11 Januari 2017

Penulis



**Fariska Dwi Ariyanto.**

## DAFTAR ISI

HALAMAN JUDUL	
LEMBAR PERSETUJUAN .....	i
LEMBAR PENGESAHAN .....	ii
LEMBAR PERNYATAAN .....	i
KATA PENGANTAR .....	ii
ABSTRAK .....	v
ABSTRACT .....	vi
LEMBAR PERSEMBAHAN .....	vii
BAB 1 PENDAHULUAN .....	1
1.1 Latar Belakang .....	1
1.2 Rumusan Masalah .....	3
1.3 Batasan Masalah .....	3
1.4 Tujuan .....	3
1.5 Metodologi .....	3
1.5.1 Identifikasi Masalah .....	4
1.5.2 Penetapan Tujuan Penelitian .....	4
1.5.3 Studi Literatur .....	4
1.5.4 Analisa Data dan Desain Sistem .....	4
1.5.5 Implementasi Sistem .....	5
1.5.6 Pengujian dan Evaluasi .....	5
1.5.7 Penyusunan Laporan .....	6
1.6 Sistematika Penulisan .....	6
BAB 2 LANDASAN TEORI .....	8
2.1 Information Retrieval .....	8
2.1.1 Pengantar Information Retrieval .....	8
2.2 Mesin Pencari( <i>Search Engine</i> ) .....	9
2.2.1 Definisi .....	9
2.3 <i>Text Mining</i> .....	10
2.3.1 <i>Case Folding</i> .....	11
2.3.2 <i>Tokenizing</i> .....	12

2.3.3	<i>Filtering / Stop Word Removal</i> .....	12
2.3.4	<i>Stemming</i> .....	13
2.4	Faktor kekurangan proses PreProcessing .....	13
2.5	Indexing dokumen.....	14
2.6	Pembobotan TF-IDF ( <i>Term Frequency – Inverse Document Frequency</i> ) 14	
2.7	Pembobotan kata dengan TW-IDF ( <i>Term Weight - Inverse Document Frequency</i> ) .....	15
2.8	<i>Cosine Similiarity</i> .....	16
2.9	Perangkingan Dokumen .....	16
2.10	Evaluasi pengujian <i>information retrieval</i> .....	17
2.10.1	<i>Evaluation of unranked retrieval sets</i> .....	17
2.10.2	<i>Evaluation of ranked retrieval results</i> .....	18
2.11	Evaluasi statistika .....	19
2.11.1	<i>Statistical significance tests</i> .....	19
2.12	Pengujian estimasi waktu respon .....	20
2.13	MySQL .....	20
BAB 3	ANALISA DAN PERANCANGAN SISTEM .....	21
3.1	Analisa Masalah .....	21
3.2	Data Penelitian.....	21
3.3	PreProcessing Data.....	24
3.4	Pembentukan Index.....	26
3.5	Pembobotan Kata TF-IDF .....	27
3.6	Pembobotan Kata TW-IDF .....	28
3.7	Perhitungan <i>Cosine Similiarity</i> .....	29
3.8	Perancangan Pengujian.....	31
3.8.1	Pengujian Perbedaan Tingkat Kecepatan Respon.....	31
3.8.2	<i>Evaluation of ranked retrieval results</i> .....	32
3.8.3	<i>Statistical significance tests</i> .....	33
BAB 4	IMPLEMENTASI DAN PENGUJIAN.....	36
4.1	Implementasi Perangkat Lunak .....	36
4.1.1	Persiapan Data .....	36
4.1.2	<i>Preprocessing Data</i> .....	36

4.1.3	Pembentukan <i>Inverted Index</i> .....	41
4.1.4	Pembobotan TF-IDF .....	42
4.1.5	Pembobotan TW-IDF .....	44
4.1.6	<i>Cosine Similiarity</i> .....	45
4.1.7	<i>Search Engine</i> .....	46
4.2	Pengujian .....	47
4.2.1	Implementasi <i>Interface</i> .....	47
4.2.2	Klasifikasi Data Hasil Pengujian.....	48
4.2.3	Metode Pengujian dan Pengolahan Data Hasil Pengujian .....	49
4.3	Kesalahan Hasil <i>Preprocessing</i> .....	60
4.4	Kesalahan Hasil Pengolahan <i>Query</i> .....	61
4.5	Evaluasi dan Analisa Hasil.....	62
BAB 5	PENUTUP .....	64
5.1	Kesimpulan .....	64
5.2	Saran.....	65
LAMPIRAN	.....	68

## DAFTAR TABEL

Tabel 3.1 Data tugas akhir .....	22
Tabel 3.2 Data <i>stoplist</i> bahasa indonesia .....	23
Tabel 3.3 Data <i>stoplist</i> bahasa inggris.....	23
Tabel 3.4 Data kata dasar .....	23
Tabel 3.5 Contoh dari data <i>inverted index</i> .....	27
Tabel 3.6 Perancangan pengujian parameter kecepatan .....	32
Tabel 3.7 Contoh perancangan pengujian relevansi TF-IDF .....	33
Tabel 3.8 Contoh perancangan pengujian relevansi TW-IDF .....	33
Tabel 3.9 Contoh perhitungan Mean Average Precision pada setiap topik .....	33
Tabel 3.10 Contoh perhitungan Wilcoxon Sign Rank Test .....	34
Tabel 4.1 Tabel z - score .....	49
Tabel 4.2 Table uji relevansi TF-IDF dengan nilai <i>average precision</i> terendah ...	50
Tabel 4.3 Tabel uji relevansi TW-IDF dengan nilai <i>average Precision</i> terendah .	50
Tabel 4.4 Tabel uji relevansi TF-IDF dengan nilai selisih terkecil dengan TW-IDF .....	51
Tabel 4.5 Tabel uji relevansi TW-IDF dengan nilai selisih terkecil dengan TF-IDF .....	51
Tabel 4.6 Tabel uji relevansi TF-IDF dengan nilai selisih terkecil dengan TW-IDF .....	52
Tabel 4.7 Tabel uji relevansi TW-IDF dengan nilai selisih terkecil dengan TF-IDF .....	52
Tabel 4.8 Tabel uji relevansi TF-IDF dengan nilai <i>average precision</i> sama.....	53
Tabel 4.9 Tabel uji relevansi TW-IDF dengan nilai <i>average precision</i> sama .....	53
Tabel 4.10 Tabel uji relevansi TF-IDF dengan nilai <i>average precision</i> tertinggi .	54
Tabel 4.11 Tabel uji relevansi TW-IDF dengan nilai <i>average precision</i> TF-IDF tertinggi.....	54
Tabel 4.12 Tabel uji relevansi TF-IDF dengan nilai <i>average precision</i> TW-IDF tertinggi .....	54
Tabel 4.13 Tabel uji relevansi TW-IDF dengan nilai <i>average precision</i> tertinggi	54
Tabel 4.14 Perhitungan <i>mean average precision</i> TF-IDF dan TW-IDF tahap pertama .....	55
Tabel 4.15 Perhitungan <i>mean average precision</i> TF-IDF dan TW-IDF tahap kedua.....	56
Tabel 4.16 Perhitungan <i>wilcoxon sign rank test</i> tahap pertama.....	57
Tabel 4.17 Perhitungan <i>wilcoxon sign rank test</i> tahap kedua .....	58
Tabel 4.18 Perhitungan estimasi waktu eksekusi pencarian dokumen .....	60

## DAFTAR GAMBAR

Gambar 2.1 Tahapan Text Mining .....	11
Gambar 2.2 Contoh tahapan <i>Case Folding</i> .....	11
Gambar 2.3 Contoh tahapan <i>tokenizing</i> .....	12
Gambar 2.4 Contoh tahapan <i>Filtering (Stop Word Removal)</i> .....	13
Gambar 2.5 Contoh tahapan <i>Stemming</i> .....	13
Gambar 2.6 Contoh pembobotan kata dan perangkian .....	16
Gambar 2.7 Grafik <i>precision / recall</i> .....	18
Gambar 2.8 Kalkulasi dari 11 poin dari interpolasi <i>average precision</i> .....	19
Gambar 3.1 Preprocessing data.....	24
Gambar 3.2 Proses <i>case folding</i> .....	24
Gambar 3.3 Proses <i>tokenizing</i> .....	24
Gambar 3.4 Proses <i>stop word removal</i> .....	25
Gambar 3.5 Proses <i>stemming</i> .....	26
Gambar 3.6 Proses pembobotan TF-IDF .....	28
Gambar 3.7 Contoh perhitungan nilai TW .....	29
Gambar 3.8 Proses pembobotan TW-IDF.....	29
Gambar 3.9 Proses perangkian dokumen dengan <i>cosine similiarity</i> .....	30
Gambar 3.10 Prototipe sistem pencarian dokumen.....	34
Gambar 3.11 Prototipe halaman login untuk kuisioner .....	35
Gambar 3.12 Prototipe pengujian relevansi .....	35
Gambar 4.1 <i>Source Code</i> Fungsi <i>Case Folding</i> .....	36
Gambar 4.2 <i>Source code</i> fungsi <i>tokenizing</i> .....	37
Gambar 4.3 Data tugas akhir sebelum proses <i>tokenizing</i> dan ( <i>case folding</i> ).....	37
Gambar 4.4 Data tugas akhir sesudah proses proses <i>tokenizing</i> dan ( <i>case folding</i> ) .....	38
Gambar 4.5 <i>Source code</i> fungsi <i>stop word removal</i> .....	38
Gambar 4.6 Data tugas akhir sebelum proses <i>stop word removal</i> .....	39
Gambar 4.7 Data tugas akhir sesudah proses <i>stop word removal</i> .....	39
Gambar 4.8 <i>Source code</i> fungsi <i>stemming</i> .....	40
Gambar 4.9 Data tugas akhir sebelum dan sesudah proses <i>stemming</i> .....	40
Gambar 4.10 <i>Source code</i> pembentuk <i>inverted index</i> .....	41
Gambar 4.11 Hasil dari pembentuk <i>inverted index</i> .....	41
Gambar 4.12 <i>Source code</i> <i>back up</i> data <i>inverted index</i> .....	41
Gambar 4.13 Hasil dari <i>back up</i> data <i>inverted index</i> .....	42
Gambar 4.14 <i>Source code</i> perhitungan nilai IDF.....	42
Gambar 4.15 <i>Source code</i> perhitungan nilai TF .....	43
Gambar 4.16 <i>Source code</i> keseluruhan perhitungan TF-IDF .....	43
Gambar 4.17 <i>Source code</i> perhitungan nilai IDF.....	44
Gambar 4.18 <i>Source code</i> menyimpan data ke dalam tabel sementara .....	44
Gambar 4.19 <i>Source code</i> perhitungan nilai TW .....	45



Gambar 4.20 Source code perhitungan nilai secara keseluruhan.....	45
Gambar 4.21 <i>Source code</i> perhitungan <i>cosine Similiarity</i> .....	46
Gambar 4.22 <i>Source code</i> implementasi <i>search engine</i> .....	46
Gambar 4.23 Tampilan login untuk volunteer sebelum mengisi kuisisioner .....	47
Gambar 4.24 Tampilan <i>search engine</i> .....	47
Gambar 4.25 Kesalahan dalam <i>preprocessing</i> data .....	61
Gambar 4.26 Kekurangan dalam pengolahan <i>query</i> .....	61



## DAFTAR FORMULA

Formula persamaan 2.1 fungsi scoring TF-IDF(1) .....	15
Formula persamaan 2.2 fungsi scoring TW-IDF(2).....	15
Formula persamaan 2.3 perhitungan <i>Cosine Similiarity</i> (3) .....	16
Formula persamaan 2.4 perhitungan <i>precision</i> (4) .....	17
Formula persamaan 2.5 perhitungan <i>recall</i> (5) .....	17
Formula persamaan 3.1 fungsi scoring TF-IDF(6) .....	27
Formula persamaan 3.2 fungsi scoring TW-IDF(7).....	28
Formula persamaan 4.1 perhitungan data sampel(8) .....	48



## DAFTAR PUSTAKA

- [1] Tonta, Yasar, 1992, *Analysis of Search Failures in Document Retrieval Systems: A Review*, University of Houston.
- [2] Rousseau, François., & Vazirgiannis, Michalis, 2013, *Graph-of-word and TW-IDF: New Approach to Ad Hoc IR*, ACM.
- [3] Sanderson, Mark., & Croft, W.Bruce, 2012, *The History of Information Retrieval Research*, IEEE.
- [4] Manning, Christopher D., Raghavan, Prabhakar., and Schütze, Hinrich., 2009, *An Introduction to Information Retrieval*, Cambridge : Cambridge University Press.
- [5] Croft, W.Bruce., Metzler, Donald., Strohman, Trevor, 2015, *Search Engine : information retrieval in practice*, Pearson Education.inc.
- [6] Knoth, Petr., Gooch, Phil, 2015, *An Introduction to Text Mining Research Papers*, Mendeley.
- [7] Asian, Jelita, 2007, *Effective Techniques for Indonesian Text Retrieval*, School of Computer Science and Information Technology, RMIT University.
- [8] Dubois, Paul, 2014, *MySQL Cookbook , Third Edition*, California : O’Rielly Media.Inc.
- [9] Cleverdon, Cyril., and Keen, Michael, 1966, *Factors Determining the Performance of Indexing Systems, Volume 2, Test Results, i ("Summary")*, Bedford.: Cranfield University.
- [10] Fitriyah, Putri, 2013, *Pengaruh Pembobotan Pada Tweet Di Mesin Pencari Menggunakan Metode TF-IDF*, Tugas Akhir Program Studi Teknik Informatika Universitas Muhammadiyah Malang.
- [11] Dai, Mashar Eka Putra., Koniyo, Moh Hidayat., & Bouty, Abd Aziz, 2014, *Temu Kembali Informasi Pencarian Dokumen Karya Ilmiah Dengan Metode Term Frequency Inverse Document Frequency (TF - IDF)*, Tugas Akhir Fakultas Teknik Program Studi Sains dan Matematika, Universitas Negeri Gorontalo.
- [12] Blanco, Roi., and Lioma, Christina, 2012, *Graph-based Term Weighting for information retrieval*, ACM.

[13 ]Survey Software Tool for Profesional Reserach | Survery Monkey.

diakses pada 20, November, 2016 dari <https://www.surveymonkey.com/mp/sample-size-calculator/>.

